eLite-HTS: a NLP tool for French HMM-based speech synthesis

Sophie Roekhaut[1], *Sandrine Brognaux*[1,2,3], *Richard Beaufort*[4], *Thierry Dutoit*[3]

[1]Cental - Université catholique de Louvain, Belgium [2]F.R.S-FNRS Research Fellow
[3]TCTS - Université de Mons, Belgium , [4]Nuance Communications, Inc., Belgium*

sophie.roekhaut@uclouvain.be, sandrine.brognaux@uclouvain.be,
richard.beaufort@nuance.com, thierry.dutoit@umons.ac.be

Abstract

This paper presents eLite-HTS, a web service which generates input files for the training and synthesis stages of a French HMM-based synthesizer using the HTS toolkit.

1. Introduction

Text-to-speech (TTS) synthesis requires a rich annotation of the text and sound at different linguistic levels (e.g. phonetics, syllables, parts of speech). Such annotations are usually automatically provided by basic language-dependent natural language processing (NLP) components such as part-of-speech (POS) taggers and grapheme-to-phoneme converters. They represent important contextual information, both for unit-selection TTS or speech synthesis with hidden Markov models (HMM). Based on these annotation layers, more complex information, like rhythmic group boundaries, can also be extracted and can offer a more complete description of suprasegmental phenomena. The Festival Speech Synthesis System [1], initially developed for unit selection synthesis, includes tools to produce such a complete linguistic analysis of the text. It has been integrated in the free HTS toolkit [2], widely used for HMM-based speech synthesis. Festival works on English and some additional languages but provides no annotation tool for French.

Few publicly available NLP tools have been proposed for French. An exception is the open source system LIA_PHON [3], which offers grapheme-to-phoneme conversion and grammatical analysis to provide a description file for MBROLA TTS [4]. This tool, however, does not include prosodic and rhythmic information as required for HMM training with HTS or for the selection of units in TTS. The IrcamAlign tool [5] proposed the generation of input HTS files for French, based on LIA_Phon analyzer, but this tool was not made available.

This paper presents eLite-HTS, a web service available at http://cental.uclouvain.be/elitehts/v1/ which generates input files for the training and synthesis stages of a French HMM-based synthesizer using the HTS toolkit. The annotation levels required for the generation of HTS files are automatically computed from text with a complete NLP tool, eLite [6, 7]. eLite is a unit selection speech synthesiser for French that includes its own full NLP service, including both basic linguistic information (e.g. phonetics, POS) and more complex rhythmic and phrasing annotation (e.g. rhythmic groups). Our web service is in a REST architectural style [8]. A fundamental principle of a REST architecture service is to keep the implementation of the server and the client independent. REST Web services can use the HTTP protocol for the communication between the server and the client. The client of the service can be implemented in any language that allows for HTTP requests (e.g. php, C, perl, Matlab). In addition to the HTS file, our web service can also produce an output in TextGrid format for the Praat application [9] dedicated to phonetic and phonological research. This web service is the first stage of a larger system that we are currently developing and which should allow for fully automatic training of French HMM speech synthesis from raw text and sound.

* The study was carried out while Richard Beaufort was still working at the CENTAL (Université catholique de Louvain, Belgium)

2. Architecture of the NLP components

This section briefly presents the different components of our NLP that will help to generate HTS description files. eLite NLP includes 3 modules (see Figure 1). A disambiguation module is made of preprocessing, morphological and syntactic analysis. A grapheme-to-phoneme conversion module is then responsible for the generation of the phonetic transcription. Finally, the prosodic module generates linguistically-driven phrasing and rhythmic information. The communication between the 3 modules is realized with a Multi-Layer Data Structure (MLDS) inspired by the structure proposed by the Festival Project. The MLDS is composed of 7 annotation layers, corresponding to different levels of segmentation of the text. *Sentences*, *tokens*, *units* and *words* are detected from the input text by the disambiguation model. *Phonemes* are generated by the grapheme-to-phoneme converter and, finally, *syllables* and *rhythmic groups* are produced by the prosodic module.

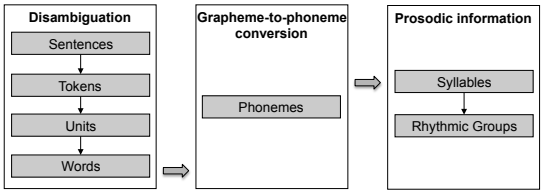

Figure 1: *Relations between the 3 modules of eLite in the MLDS*

First, in the disambiguation module, a preprocessing of the text is applied, based on a set of manually-tuned rewrite rules. It identifies sentence boundaries and detects some special tokens : URLs, phone numbers, dates, times, currencies and units of measurement. It produces the *sentence* and *token* units. Then, it carries out, sentence by sentence, a morphological analysis and a contextual disambiguation. The morphological analysis

performs, at the *word* level, a lexicon look-up to determine the set of possible grammatical categories for each word, given its token. It then creates the *unit* level that groups together some *words* like compound nouns (e.g. "pomme de terre", "potato"). A contextual disambiguation at the *unit* level selects then the most likely grammatical category from the set of propositions, based on a statistical language model [10].

The grapheme-to-phoneme conversion module first produces the phonetic transcription of each *word* unit, with *phoneme* units, regardless its context. Our phonetisation is based on an ID3 decision tree learned from a phonetic dictionary [11]. Note that the dictionary is automatically accentuated. For non-clitic words, a primary stress is put on the last vowel while a secondary stress is put on the first. Previously extracted contextual information is then exploited to deal with phonetic variations at word boundaries. The most frequent phenomena that appear in French are the liaisons between two words and the deletion and insertion of schwa for the lubrication of speech.

The prosodic module provides linguistic information that contributes in generating a more natural prosody in HMM-based or unit selection speech synthesis. First, *syllables* are determined with a set of rules based on the possible grouping of *phonemes*, according to articulatory criteria. Each *syllable* is then associated with the stress status of its corresponding nucleus. In a second stage, grammatical information is used to detect the boundaries of *rhythmic groups*, which should correspond to breath groups ending in a boundary tone. The rhythmic groups are detected with a chinks and chunks algorithm [12]. Figure 2 gives an example of sentence in TextGrid format as output by eLite NLP.

<table>
<tr><td>Words</td><td colspan="3">Les</td><td colspan="4">oiseaux</td><td colspan="3">mouches</td><td colspan="3">chantent</td><td colspan="3">sur</td><td colspan="2">la</td><td colspan="4">branche</td><td>.</td></tr>
<tr><td>Phonemes</td><td>l</td><td>e</td><td>z</td><td>w</td><td>a</td><td>z</td><td>o</td><td>m</td><td>u</td><td>S</td><td>S</td><td>a~</td><td>t</td><td>s</td><td>y</td><td>R</td><td>l</td><td>a</td><td>b</td><td>R</td><td>a~</td><td>S</td><td>_</td></tr>
<tr><td>Syllables</td><td colspan="2">l e</td><td colspan="3">z w a</td><td colspan="2">z o</td><td colspan="3">m u S</td><td colspan="3">S a~ t</td><td colspan="3">s y R</td><td colspan="2">l a</td><td colspan="4">b R a~ S</td><td>_</td></tr>
<tr><td>Units</td><td colspan="3">DET</td><td colspan="7">NOUN</td><td colspan="3">VERB</td><td colspan="3">PREP</td><td colspan="2">DET</td><td colspan="4">NOUN</td><td>PUNCT</td></tr>
<tr><td>Rythm groups</td><td colspan="10">RG</td><td colspan="3">RG</td><td colspan="9">RG</td><td>PUNCT</td></tr>
<tr><td>Sentences</td><td colspan="23">DECLARATIVE</td></tr>
</table>

Figure 2: *Example of NLP analysis for sentence in TextGrid format "Les oiseaux-mouches chantent sur la branche (The hummingbirds are singing on the branch)"*

3. Generation of the HTS file

The minimal linguistic unit used in the HTS description files is the phoneme. For each phoneme, 53 characteristics are described. The other linguistic levels used by HTS are (from the low- to high-level) : syllables, words, phrases and utterances. Each unit is described according to its context (left and right), its position in relation to the other linguistic units of higher level and the number of units of the lower level it contains. Some specific features are added to the different levels : stresses for syllables, POS for words and ToBI endtones for phrases [13]. The internal structure of eLite MLDS provides most of the data required for the generation of HTS files.

Only ToBI tones are not automatically generated by eLite-HTS. While this annotation system has been widely adopted by the English-speaking community, several prosody annotation protocols are commonly used for French prosody. The web service offers the possibility to introduce prosody information, manually or with external scripts. To use this functionality, a TextGrid should first be generated by eLite-HTS. A specific annotation layer, aligned with the syllable tier, can then be filled with prosody information. eLite-HTS then allows for the generation of HTS files considering this new annotation level. Such an integration of prosodic information was used for instance in [14, 15] for the synthesis of sports commentaries.

Two types of output files can be produced, corresponding to the training or synthesis stages. Conversely to synthesis description files, the training files output by eLite-HTS additionally contain the time boundary of each phoneme in the sound files. To obtain these boundaries, the TextGrid output by the web service can be further processed by automatic aligners like-Train&Align [16] or EasyAlign [17]. The TextGrid can then be provided once more to elite-HTS to produce training HTS files.

4. Future works

Perspectives include the connection of the web service with Train&Align, for the automatic alignment of sound files before training, and HTS modules to fully generate French HMM-based speech synthesis. Later versions should also allow for more personalization of the different NLP modules.

5. References

[1] A. Black, P. Taylor, R. Caley, R. Clark, K. Richmond, S. King, V. Strom, and H. Zen, "The festival speech synthesis system, version 1.4. 2," *Unpublished document available via http://www. cstr. ed. ac. uk/projects/festival. html*, 2001.

[2] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. Black, and K. Tokuda, "The HMM-based speech synthesis system (HTS) version 2.0," in *SSW6*, 2007, pp. 294–299.

[3] F. Béchet, "LIA PHON: un système complet de phonétisation de textes," *Traitement automatique des langues*, vol. 42, no. 1, pp. 47–67, 2001.

[4] T. Dutoit, V. Pagel, N. Pierret, F. Bataille, and O. Van der Vrecken, "The MBROLA project: Towards a set of high quality speech synthesizers free of use for non commercial purposes," in *ICSLP*, 1996, pp. 1393–1396.

[5] C. Veaux, G. Beller, D. Schwarz, and X. Rodet, "Ircamcorpustools: an extensible platform for speech corpora exploitation," in *LREC*, 2008, pp. 3398–3401.

[6] R. Beaufort and A. Ruelle, "eLite: système de synthèse de la parole à orientation linguistique," *JEP*, pp. 509–512, 2006.

[7] V. Colotte and R. Beaufort, "Linguistic features weighting for a text-to-speech system without prosody model," in *Interspeech*, 2005, pp. 2549–2552.

[8] L. Richardson and S. Ruby, *RESTful web services*. O'Reilly, 2008.

[9] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot international*, vol. 5, no. 9/10, pp. 341–345, 2002.

[10] R. Beaufort, T. Dutoit, and V. Pagel, "Analyse syntaxique du français. pondération par trigrammes lissés et classes d'ambiguïtés lexicales," *JEP*, pp. 133–136, 2002.

[11] V. Pagel, K. Lenzo, and A. Black, "Letter to sound rules for accented lexicon compression," in *ICSLP*, 1998, pp. 252–255.

[12] M. Y. Liberman and K. W. Church, "Text analysis and word pronunciation in text-to-speech synthesis," *Advances in speech signal processing*, pp. 791–831, 1992.

[13] C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg, "Tobi: A standard for labeling english prosody." in *ICSLP*, 1992, pp. 12–16.

[14] B. Picart, S. Brognaux, and T. Drugman, "HMM-based speech synthesis of live sports commentaries: Integration of a two-layer prosody annotation." in *SSW8*, 2013.

[15] S. Brognaux, B. Picart, and T. Drugman, "A new prosody annotation protocol for live sports commentaries," in *Interspeech*, 2013.

[16] S. Brognaux, S. Roekhaut, T. Drugman, and R. Beaufort, "Train&Align: A new online tool for automatic phonetic alignment," in *SLT*, 2012.

[17] J. P. Goldman, "Easyalign: An automatic phonetic alignment tool under praat." in *Interspeech*, 2011, pp. 3233–3236.